

ADVANCES AND CHALLENGES IN FOUNDATION AGENTS

FROM BRAIN-INSPIRED INTELLIGENCE TO EVOLUTIONARY, COLLABORATIVE, AND SAFE SYSTEMS

Bang Liu^{2,3,20*†}, Xinfeng Li^{4*}, Jiayi Zhang^{1,10*}, Jinlin Wang^{1*}, Tanjin He^{5*}, Sirui Hong^{1*},
 Hongzhang Liu^{6*}, Shaokun Zhang^{7*}, Kaitao Song^{8*}, Kunlun Zhu^{9*}, Yuheng Cheng^{1*},
 Suyuchen Wang^{2,3*}, Xiaoqiang Wang^{2,3*}, Yuyu Luo^{10*}, Haibo Jin^{9*}, Peiyan Zhang¹⁰, Ollie Liu¹¹,
 Jiaqi Chen¹, Huan Zhang^{2,3}, Zhaoyang Yu¹, Haochen Shi^{2,3}, Boyan Li¹⁰, Dekun Wu^{2,3}, Fengwei Teng¹,
 Xiaojun Jia⁴, Jiawei Xu¹, Jinyu Xiang¹, Yizhang Lin¹, Tianming Liu¹⁴, Tongliang Liu⁶,
 Yu Su¹⁵, Huan Sun¹⁵, Glen Berseth^{2,3,20}, Jianyun Nie², Ian Foster⁵, Logan Ward⁵, Qingyun Wu⁷,
 Yu Gu¹⁵, Mingchen Zhuge¹⁶, Xiangru Tang¹², Haohan Wang⁹, Jiaxuan You⁹, Chi Wang¹⁹,
 Jian Pei^{17†}, Qiang Yang^{10,18†}, Xiaoliang Qi^{13†}, Chenglin Wu^{1*†}

¹MetaGPT, ²Université de Montréal, ³Mila - Quebec AI Institute, ⁴Nanyang Technological University,
⁵Argonne National Laboratory, ⁶University of Sydney, ⁷Penn State University, ⁸Microsoft Research Asia,
⁹University of Illinois at Urbana-Champaign, ¹⁰The Hong Kong University of Science and Technology,
¹¹University of Southern California, ¹²Yale University, ¹³Stanford University, ¹⁴University of Georgia,
¹⁵The Ohio State University, ¹⁶King Abdullah University of Science and Technology, ¹⁷Duke University,
¹⁸The Hong Kong Polytechnic University, ¹⁹Google DeepMind, ²⁰Canada CIFAR AI Chair

ABSTRACT

The advent of large language models (LLMs) has catalyzed a transformative shift in artificial intelligence, paving the way for advanced intelligent agents capable of sophisticated reasoning, robust perception, and versatile action across diverse domains. As these agents increasingly drive AI research and practical applications, their design, evaluation, and continuous improvement present intricate, multifaceted challenges. This survey provides a comprehensive overview, framing intelligent agents within a modular, brain-inspired architecture that integrates principles from cognitive science, neuroscience, and computational research. We structure our exploration into four interconnected parts. First, we delve into the **modular foundation of intelligent agents**, systematically mapping their cognitive, perceptual, and operational modules onto analogous human brain functionalities, and elucidating core components such as memory, world modeling, reward processing, and emotion-like systems. Second, we discuss **self-enhancement and adaptive evolution mechanisms**, exploring how agents autonomously refine their capabilities, adapt to dynamic environments, and achieve continual learning through automated optimization paradigms, including emerging AutoML and LLM-driven optimization strategies. Third, we examine **collaborative and evolutionary multi-agent systems**, investigating the collective intelligence emerging from agent interactions, cooperation, and societal structures, highlighting parallels to human social dynamics. Finally, we address the critical imperative of **building safe, secure, and beneficial AI systems**, emphasizing intrinsic and extrinsic security threats, ethical alignment, robustness, and practical mitigation strategies necessary for trustworthy real-world deployment. By synthesizing modular AI architectures with insights from different disciplines, this survey identifies key research gaps, challenges, and opportunities, encouraging innovations that harmonize technological advancement with meaningful societal benefit. The project's Github link is: <https://github.com/FoundationAgents/awesome-foundation-agents>.

*Major Contribution. Work in progress.

†Corresponding authors: Bang Liu (bang.liu@umontreal.ca), Jian Pei (j.pei@duke.edu), Qiang Yang (qyang@cse.ust.hk), Xiaoliang Qi (xlqi@stanford.edu), Chenglin Wu (alexanderwu@deepwisdom.ai)

Preface

Large language models (LLMs) have revolutionized artificial intelligence (AI) by demonstrating unprecedented capabilities in natural language and multimodal understanding, as well as reasoning and generation. These models are trained on vast datasets, and they exhibit emergent abilities such as reasoning, in-context learning, and even rudimentary planning. While these models represent a major step forward in realizing intelligent machines, they themselves do not yet fully embody all the capabilities of an intelligent being. Since the early days of artificial intelligence, AI researchers have long been on a quest for a truly “intelligent” system that can learn, plan, reason, sense, communicate, act, remember, and demonstrate various human-like abilities and agility. These beings, known as intelligent agents, should be able to think both long-term and short-term, perform complex actions, and interact with humans and other agents. LLMs are an important step towards realizing intelligent agents, but we are not there yet.

This manuscript provides a comprehensive overview of the current state of the art of LLM-based intelligent agents. In the past, there have been numerous research papers and books on intelligent agents, as well as a flurry of books on LLMs. However, there has scarcely been comprehensive coverage of both. While LLMs can achieve significant capabilities required by agents, they only provide the foundations upon which further functionalities must be built. For example, while LLMs can help generate plans such as travel plans, they cannot yet generate fully complex plans for complex and professional tasks, nor can they maintain long-term memories without hallucination. Furthermore, their ability to perform real-world actions autonomously remains limited. We can view LLMs as engines, with agents being the cars, boats, and airplanes built using these engines. In this view, we naturally seek to move forward in designing and constructing fully functioning intelligent agents by making full use of the capabilities provided by LLMs.

In this engine-vehicle analogy of the interplay between LLMs and agents, we naturally ask: How much of the capabilities of intelligent agents can current LLM technologies provide? What are the functions that cannot yet be realized based on current LLM technologies? Beyond LLMs, what more needs to be done to have a fully intelligent agent capable of autonomous action and interaction in the physical world? What are the challenges for fully integrated LLM-based agents? What additional developments are required for capable, communicative agents that effectively collaborate with humans? What are the areas that represent low-hanging fruits for LLM-based agents? What implications will there be for society once we have fully intelligent LLM-based agents, and how should we prepare for this future?

These questions transcend not only the engineering practice of extending current LLMs and agents but also raise potential future research directions. We have assembled frontier researchers from AI, spanning from LLM development to agent design, to comprehensively address these questions. The book consists of four parts. The first part presents an exposition of the requirements for individual agents, comparing their capabilities with those of humans, including perception and action abilities. The second part explores agents’ evolution capabilities and their implications on intelligent tools such as workflow management systems. The third part discusses societies of agents, emphasizing their collaborative and collective action capabilities, and the fourth part addresses ethical and societal aspects, including agent safety and responsibilities.

This book is intended for researchers, students, policymakers, and practitioners alike. The audience includes non-AI readers curious about AI, LLMs, and agents, as well as individuals interested in future societies where humans co-exist with AI. Readers may range from undergraduate and graduate students to researchers and industry practitioners. The book aims not only to provide answers to readers’ questions about AI and agents but also to inspire them to ask new questions. Ultimately, we hope to motivate more people to join our endeavor in exploring this fertile research ground.

Contents

1	Introduction	12
1.1	The Rise and Development of AI Agents	12
1.2	A Parallel Comparison between Human Brain and AI Agents	13
1.2.1	Brain Functionality by Region and AI Parallels	14
1.3	A Modular and Brain-Inspired AI Agent Framework	16
1.3.1	Core Concepts and Notations in the Agent Loop	18
1.3.2	Biological Inspirations	21
1.3.3	Connections to Existing Theories	21
1.4	Navigating This Survey	22
I	Core Components of Intelligent Agents	24
2	Cognition	25
2.1	Learning	25
2.1.1	Learning Space	27
2.1.2	Learning Objective	29
2.2	Reasoning	31
2.2.1	Structured Reasoning	32
2.2.2	Unstructured Reasoning	34
2.2.3	Planning	36
3	Memory	39
3.1	Overview of Human Memory	39
3.1.1	Types of Human Memory	39
3.1.2	Models of Human Memory	41
3.2	From Human Memory to Agent Memory	42
3.3	Representation of Agent Memory	44
3.3.1	Sensory Memory	44
3.3.2	Short-Term Memory	46
3.3.3	Long-Term Memory	46
3.4	The Memory Lifecycle	47

3.4.1	Memory Acquisition	47
3.4.2	Memory Encoding	48
3.4.3	Memory Derivation	49
3.4.4	Memory Retrieval and Matching	50
3.4.5	Neural Memory Networks	51
3.4.6	Memory Utilization	52
3.5	Summary and Discussion	53
4	World Model	54
4.1	The Human World Model	55
4.2	Translating Human World Models to AI	55
4.3	Paradigms of AI World Models	56
4.3.1	Overview of World Model Paradigms	56
4.3.2	Implicit Paradigm	57
4.3.3	Explicit Paradigm	57
4.3.4	Simulator-Based Paradigm	58
4.3.5	Hybrid and Instruction-Driven Paradigms	58
4.3.6	Comparative Summary of Paradigms	58
4.4	Relationships to Other Modules	58
4.4.1	Memory and the World Model	59
4.4.2	Perception and the World Model	60
4.4.3	Action and the World Model	60
4.4.4	Cross-Module Integration	61
4.5	Summary and Discussion	61
5	Reward	63
5.1	The Human Reward Pathway	64
5.2	From Human Rewards to Agent Rewards	65
5.3	AI Reward Paradigms	65
5.3.1	Definitions and Overview	65
5.3.2	Extrinsic Rewards	67
5.3.3	Intrinsic Rewards	67
5.3.4	Hybrid Rewards	68
5.3.5	Hierarchical Rewards	68
5.4	Summary and Discussion	69
5.4.1	Interaction with Other Modules	69
5.4.2	Challenges and Directions	69
6	Emotion Modeling	71
6.1	Psychological Foundations of Emotion	71
6.2	Incorporating Emotions in AI Agents	74

6.3	Understanding Human Emotions through AI	74
6.4	Analyzing AI Emotions and Personality	74
6.5	Manipulating AI Emotional Responses	75
6.6	Summary and Discussion	75
7	Perception	77
7.1	Human versus AI Perception	77
7.2	Types of Perception Representation	79
7.2.1	Unimodal Models	79
7.2.2	Cross-modal Models	80
7.2.3	Multimodal Models	81
7.3	Optimizing Perception Systems	83
7.3.1	Model-Level Enhancements	83
7.3.2	System-Level Optimizations	84
7.3.3	External Feedback and Control	84
7.4	Perception Applications	84
7.5	Summary and Discussion	85
8	Action Systems	86
8.1	The Human Action System	86
8.2	From Human Action to Agentic Action	87
8.3	Paradigms of Agentic Action System	88
8.3.1	Action Space Paradigm	88
8.3.2	Action Learning Paradigm	91
8.3.3	Tool-Based Action Paradigm	93
8.4	Action and Perception: “Outside-In” or “Inside-out”	95
8.5	Summary and Discussion	97
II	Self-Evolution in Intelligent Agents	100
9	Optimization Spaces and Dimensions for Self-evolution	103
9.1	Overview of Agent Optimization	103
9.2	Prompt Optimization	103
9.2.1	Evaluation Functions	104
9.2.2	Optimization Functions	104
9.2.3	Evaluation Metrics	105
9.3	Workflow Optimization	105
9.3.1	Workflow Formulation	105
9.3.2	Optimizing Workflow Edges	106
9.3.3	Optimizing Workflow Nodes	106

9.4	Tool Optimization	107
9.4.1	Learning to Use Tools	107
9.4.2	Creation of New Tools	107
9.4.3	Evaluation of Tool Effectiveness	108
9.5	Towards Autonomous Agent Optimization	110
10	Large Language Models as Optimizers	111
10.1	Optimization Paradigms	111
10.2	Iterative Approaches to LLM Optimization	111
10.3	Optimization Hyperparameters	114
10.4	Optimization across Depth and Time	114
10.5	A Theoretical Perspective	115
11	Online and Offline Agent Self-Improvement	116
11.1	Online Agent Self-Improvement	116
11.2	Offline Agent Self-Improvement	117
11.3	Comparison of Online and Offline Improvement	118
11.4	Hybrid Approaches	118
12	Scientific Discovery and Intelligent Evolution	120
12.1	Agent’s Intelligence for Scientific Knowledge Discovery	120
12.1.1	KL Divergence-based Intelligence Measure	120
12.1.2	Statistical Nature of Intelligence Growth	122
12.1.3	Intelligence Evolution Strategies	123
12.2	Agent-Knowledge Interactions	123
12.2.1	Hypothesis Generation and Testing	124
12.2.2	Protocol Planning and Tool Innovation	126
12.2.3	Data Analysis and Implication Derivation	126
12.3	Technological Readiness and Challenges	127
12.3.1	Real-World Interaction Challenges	127
12.3.2	Complex Reasoning Challenges	128
12.3.3	Challenges in Integrating Prior Knowledge	129
III	Collaborative and Evolutionary Intelligent Systems	130
13	Design of Multi-Agent Systems	133
13.1	Strategic Learning: Cooperation vs. Competition	133
13.2	Modeling Real-World Dynamics	134
13.3	Collaborative Task Solving with Workflow Generation	135
13.4	Composing AI Agent Teams	135
13.5	Agent Interaction Protocols	137

13.5.1	Message Types	137
13.5.2	Communication Interface	138
13.5.3	Next-Generation Communication Protocols	138
14	Communication Topology	141
14.1	System Topologies	141
14.1.1	Static Topologies	141
14.1.2	Dynamic and Adaptive Topologies	142
14.2	Scalability Considerations	144
15	Collaboration Paradigms and Collaborative Mechanisms	146
15.1	Agent-Agent collaboration	146
15.2	Human-AI Collaboration	149
15.3	Collaborative Decision-Making	150
16	Collective Intelligence and Adaptation	152
16.1	Collective Intelligence	152
16.2	Individual Adaptability	153
17	Evaluating Multi-Agent Systems	155
17.1	Benchmarks for Specific Reasoning Tasks	155
17.2	Challenge and Future Work	159
IV	Building Safe and Beneficial AI Agents	160
18	Agent Intrinsic Safety: Threats on AI Brain	163
18.1	Safety Vulnerabilities of LLMs	163
18.1.1	Jailbreak Attacks	163
18.1.2	Prompt Injection Attacks	166
18.1.3	Hallucination Risks	167
18.1.4	Misalignment Issues	169
18.1.5	Poisoning Attacks	170
18.2	Privacy Concerns	172
18.2.1	Inference of Training Data	172
18.2.2	Inference of Interaction Data	173
18.2.3	Privacy Threats Mitigation	174
18.3	Summary and Discussion	175
19	Agent Intrinsic Safety: Threats on Non-Brain Modules	176
19.1	Perception Safety Threats	176
19.1.1	Adversarial Attacks on Perception	176

19.1.2	Misperception Issues	177
19.2	Action Safety Threats	178
19.2.1	Supply Chain Attacks	178
19.2.2	Risks in Tool Usage	179
20	Agent Extrinsic Safety: Interaction Risks	180
20.1	Agent-Memory Interaction Threats	180
20.2	Agent-Environment Interaction Threats	180
20.3	Agent-Agent Interaction Threats	182
20.4	Summary and Discussion	182
21	Superalignment and Safety Scaling Law in AI Agents	184
21.1	Superalignment: Goal-Driven Alignment for AI Agents	184
21.1.1	Composite Objective Functions in Superalignment	184
21.1.2	Overcoming the Limitations of RLHF with Superalignment	185
21.1.3	Empirical Evidence Supporting Superalignment	185
21.1.4	Challenges and Future Directions	185
21.2	Safety Scaling Law in AI Agents	186
21.2.1	Current landscape: balancing model safety and performance	186
21.2.2	Enhancing safety: preference alignment and controllable design	187
21.2.3	Future directions and strategies: the AI-45° rule and risk management	187
22	Concluding Remarks and Future Outlook	189